電子情報通信学会論文誌 '95/2 Vol. J78-D-II No. 2

周辺分布特徴を用いた数式構造認識

正 員 岡本 正行†
非会員 トワキョンド ムサフィリ ハシム†
Mathematical Expression Recognition by Projection Profile Characteristics
Masayuki OKAMOTO†, *Member and* Hashim Msafiri TWAAKYONDO†, *Nonmember*

† 信州大学工学部情報工学科, 長野市
Faculty of Engineering, Shinshu University, Nagano-shi, 380 Japan

あらまし 印刷文書中の数式を読み取るために，水平・垂直方向の周辺分布による再帰的分割により，数式の大まかな構造を取り出すと共に，これらの構造を表現する木構造を走査して，部分式の構造を認識する手法について述べている．多様な構造をもつ数式画像で認識実験を行い，本手法の有効性を確認している．

キーワード 数式認識，文書解析，文書認識

1. まえがき

科学技術文献等の印刷文書を読み取り，電子ファイル化する際には，文字領域だけでなく数式や表も機械可読形式に変換しておくことが望ましい．本論文では，数式部分の読取りを行うための高速で簡易な手法を提案している．

従来，数式構造を認識する手法として，構文解析を用いる手法や，記号同士の相対的な位置関係を構文上の知識を用いて解析する手法等が提案されているが，極めて限定された範囲の数式しか認識できていない[(1)~(3)]．本論文では，文書画像の領域分割の一手法として用いられている，「周辺分布による再帰的分割」と同様な考え方に基づいた数式の構造認識手法を提案している[(4)~(6)]．数式全体はいくつかの構成要素，つまり記号や部分式が水平に配置されたものと考えられるため，まず黒画素の垂直分布が0の部分で数式を水平方向に分割する．次に分割された各領域で水平方向の周辺分布を計算し，それぞれの領域を垂直方向に分割する．この処理は，縦方向に接続関係をもつ数式構造を抽出するためである．以後これらの処理は，周辺分布による分割が不可能になるまで再帰的に繰り返される．また，これらの分割により抽出された数式の階層構造は，木構造上の水平・垂直リンクにより表現される．しかしながらこの処理では，数式中の大まかな構造は解析されるが，分数式，添字式，行列等は正しく認識されない．これらの構造は記号認識を行った後で，木構造を再走査することにより認識される．

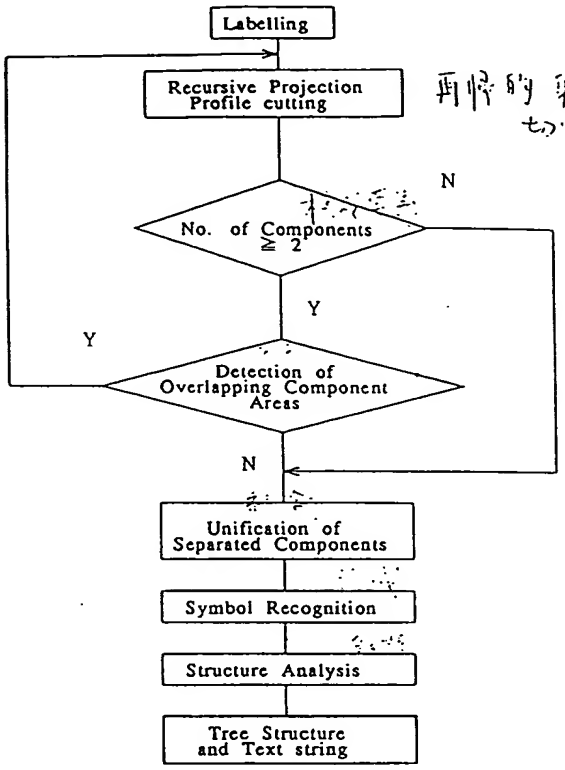


図1 数式の認識手続き
Fig. 1 Flowchart of recognition procedure.

本手法は構文解析に基づく手法に比べて，数式の生成規則を記述する必要がなく，容易に広範囲の数式に対応できるばかりでなく，高速に処理することが可能である．また，認識された数式はその構造と1対1に対応する木構造で表現されるため，TEXのような適当なフォーマッタを用いて，もとの数式が再現可能である．

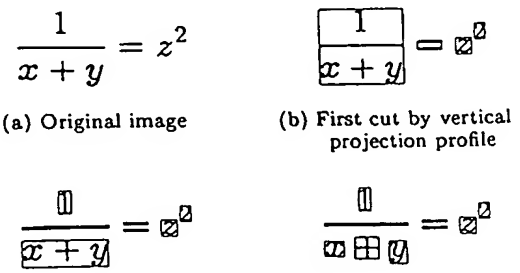
2. 数式の認識手続き

図1に処理手順を示す．本手法は，(1)周辺分布による構造解析，(2)分離記号の統合と記号認識，(3)木構造の再走査による構造解析の三つの主要部分から構成されており，以下にその詳細を述べる．

2.1 周辺分布による構造解析

(1) 周辺分布による分割

与えられた数式領域に対して，黒画素の垂直方向周辺分布を計算し，分布が0の所で分割する(図2の(b))．次に各分割された各領域に対して，水平方向の周辺分布を計算し同様に分割を行う(図2の(c))．これらの処理を分割が不可能になるまで再帰的に繰り返す(図2の(d))．このときの各分割領域は数式の階層構造を表しており，図3に示すような木構造上の水平・垂



(a) Original image (b) First cut by vertical projection profile

(c) Second cut by horizontal projection profile (d) Result by recursive projection profile

図2 周辺分布による再帰的分割
Fig. 2 Partitioning by recursive projection profile.

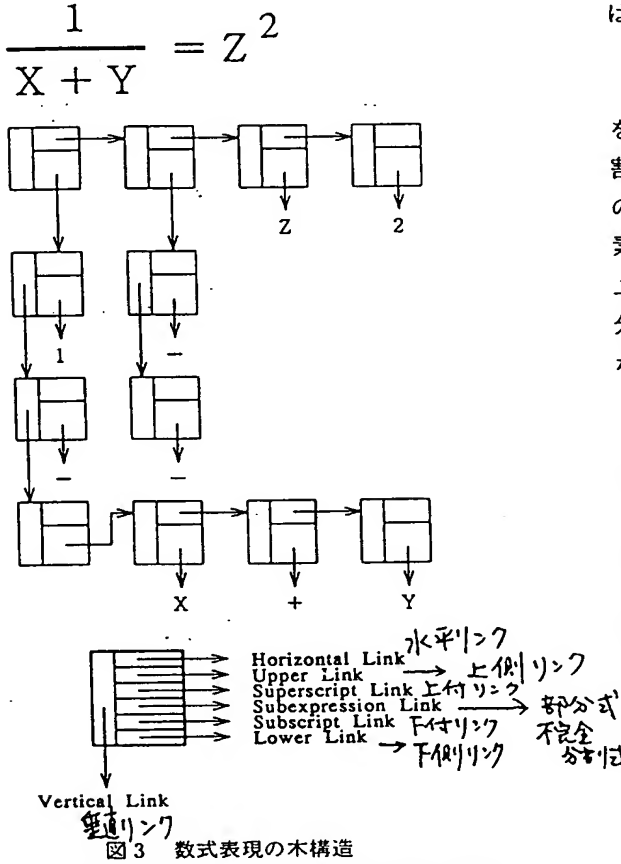
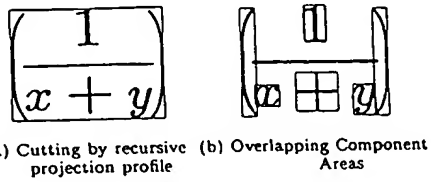


図3 数式表現の木構造
Fig. 3 Tree structure of a mathematical expression.

直方向のリンクで表現される．但しこの分割だけでは，構造を正しく解析できない数式が存在する．例えば図3では，z^2 の "2" は文字 "z" と水平方向リンクで接続されており，べき乗としては認識されていない．また右辺の分数は，分子の "1"，分数記号，分母の "$x+y$" が垂直方向リンクだけで接続されており，分数記号を中心として，上に分子，下に分母が接続された構造とはなっていない．これらの構造は，後に述べる木構造の再走査による構造解析処理で修正される．また "=" 記号は周辺分布による分割の時点では，上下に分離した記号として取り出されている．このような分離記号は，以下に述べる記号統合処理で統合される．



(a) Cutting by recursive projection profile (b) Overlapping Component Areas

図4 重なり領域の検出
Fig. 4 Detection of overlapping component areas.

（2） 重なり領域の検出

上記(1)の処理では，図4のように記号同士は重なりをもたなくても，周辺分布が0となる場所がなくて分割が行われない場合がある．これに対処するため，(1)の分割で得られた各領域に含まれているラベル数(黒画素連結領域数)をカウントし，2以上ならば左右または上下から一つずつ記号を取り除き，周辺分布による再分割を試みる．この処理は，領域に含まれるラベル数が1になるまで繰り返される．

2.2 分離記号の統合と記号認識

数式では水平あるいは垂直方向に分離した構成要素をもつ記号が多く現れる．これらの分離記号を統合することは，数式中で使用されている記号のサイズが推定できれば比較的容易であるが，一般的に数式では多くの異なるサイズの記号が用いられるため，サイズの推定による統合は困難である．このため本手法では，分離記号の構成要素を個々に認識した後，それらの統合可能性を検査する手法を採用している．この検査では，あらかじめ分離記号の構成要素ごとに，水平および垂直方向に統合可能な構成要素を示す表[6]を用いている．木構造上で水平および垂直方向リンクをもつ記号同士は，相対的大きさおよび距離と，この統合可能性を示す表から統合可能か否かが判断される．

2.3 木構造の再走査による構造解析

数式の大まかな構造は，2.1の周辺分布による再帰的分割処理で取り出されているが，添字式，分数式，上下限式等をもつ総和式，行列等は正しく構造が認識されていない．ここでは，木構造を走査してこれらの構造の認識を行い，木構造上のリンクを修正する．

2.3.1 添 字 式

ここでは，上付きまたは下付き添字式の認識を行う．

説明の都合上，まず木構造のノード間の隣接関係を定義する．あるノードに対して，そのノードと水平リンクで直接接続されている単一ノードまたは部分木のノードを水平隣接ノードと呼ぶ．添字式は，垂直リンク数が2以下の水平隣接ノードに対して，以下に述べる手順で認識される．ここで，垂直リンク数を2以下としたのは，上下の添字が同時に付く場合を考慮したためで，縦方向に印刷された分数式等が添字式となる場合は考慮していない．添字式として認識された場合，これまでの水平リンクが上付きまたは下付きのリンクに変えられ，水平リンクで接続された次のノードが水平隣接ノードとなり，更に添字式の検査が行われる．これにより，添字式の入れ子構造も同様の処理が可能となっている．

添字式の認識は，記号同士の縦方向の相対位置の比較により行っている．しかしながら，文字にはアセンディングやディセンディング文字が存在するため，単に文字方形の中心位置を比較することはできない．このため，まず文字の正規化中心を定義する．

（1） 正規化中心

正規化中心（以下 nc と記す）は，同一ベースライン上に印刷されたアセンディング，ディセンディング，および通常文字がそれぞれ同じ高さの中心位置をもつように，各文字方形の中心位置を修正したものであり，次式で定義される（図5の(a)）．

$$nc=\begin{cases} c-h/6 & \text{アセンディング文字} \\ c+h/6 & \text{ディセンディング文字} \\ c & \text{通常文字} \end{cases}$$

ここで，c，h はそれぞれ文字方形の中心位置，高さである．

（2） 添字式の判定

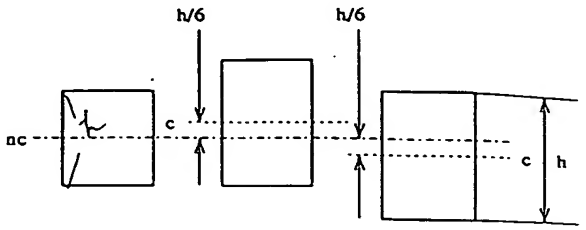
水平隣接ノード同士の記号に対して，図5の(b)に示すように，以下の条件で上付き，下付きの判断を行う．このとき，前者のノードに上付きまたは下付きリンクで接続された記号がある場合には，その記号と比較する．

$c^2_{nc}>c^1_{nc}$ and $h/6<(c^2_{nc}-c^1_{nc})<2h/3$：上付き

$c^1_{nc}>c^3_{nc}$ and $h/6<(c^1_{nc}-c^3_{nc})<2h/3$：下付き

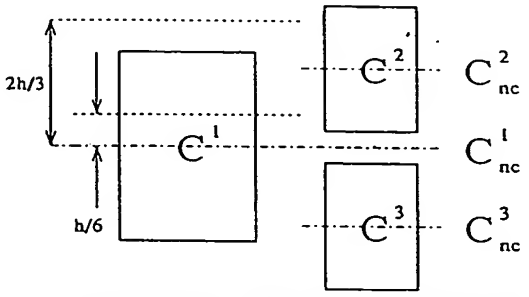
2.3.2 行　　列

周辺分布による再帰的な分割では，水平方向の分割が最初に試みられるため，行列や行列式ではいくつかの列が横方向に並んだ形の木構造が得られる．しかしながらこの木構造では，単に行列を表す括弧が列と共に水平方向に並んだ構造を表しているにすぎない．また列を中心とした表現は，TEXのような行を中心とした表現に変換しがたい．ここでは，ある程度以上の大きさの括弧ペアが検出され，それらが行列を構成している可能性がある場合，括弧に囲まれた領域について改めて周辺分布による分割を適用し，行を中心とした行列表現を得ている．行列の認識は以下の手順で行われ，木構造が変更される．



h : height of character, c : center of bounding box, nc : normalized center

(a) Definition of normalized center.



(b) Condition of a subscript or superscript

図5　添字式の認識

Fig. 5　Recognition of a subscript or superscript.

（1） 数式領域中のすべての文字方形（記号も含む）の縦幅から平均縦幅を求め，平均縦幅の3倍以上の括弧（"("，"["，"|"）のペアを見つける．

（2） 括弧を除いた領域に対して，水平方向の黒画素周辺分布により各行の縦幅を計算し，すべての行の縦幅がほぼ等しく（本論文では1.2倍以内としている）かつ行が2行以上存在するか調べる．この条件を満たしていれば行列構造が存在するものとして，以下の処理を行う．

（3） 括弧を除いた領域に対して，2.1の周辺分布による再帰的分割を，水平方向の周辺分布による分割を先にして適用する．

2.3.3 分数式および上下限式

周辺分布による再帰的分割では，分数式や上下限式をもつ多くの記号式（"Σ"，"Π"，"∪" 等）は，各部分式が単に垂直方向に並んだ構造となっており，木構造

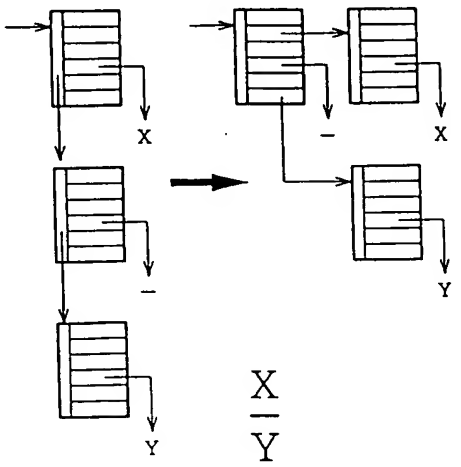


図 6　木構造の修正
Fig. 6　Correction of a tree structure.

は正しい数式構造を表現していない．このような数式に対しては，以下の手順により木構造の修正が行われる．

（1） 木構造を走査することにより，分数記号や上下限式をもつ可能性のある記号を見つける．

（2） 図 6 に示すように，この記号のノードが二つ(分数記号の場合は三つ)以上の垂直リンクにより接続された部分木のノードかどうか調べる．もし該当していれば，この記号のノードが水平隣接ノードとなり，かつこれまで垂直リンクで接続されていたノードは，上下リンクでこのノードに接続されるように変更する．

3. 実験結果および考察

いくつかの論文誌からさまざまな構造をもつ数式を抽出すると共に，任意の数式を得るために TEX の出力結果を用いて認識実験を行った．数式画像はスキャナにより 320 dpi で入力し，処理には Sun 4/2 を用いて

$$g(x,y)=\frac{1}{2\pi\sigma_x\sigma_y}\exp\left\{-\frac{1}{2}\left[\left(\frac{x}{\sigma_x}\right)^2+\left(\frac{y}{\sigma_y}\right)^2\right]\right\},$$

$$\sigma_n(t)=\frac{1}{2\pi}\int_{-\pi}^{\pi}f(x)K_n(t-x)dx,$$

$$R(p,q)=\max\left[0,\frac{1+p_sp+q_sq}{\sqrt{1+p^2+q^2}\sqrt{1+p_s^2+q_s^2}}\right].$$

$$\frac{\alpha}{\left\{2-\alpha+\frac{(1-r)\alpha}{L}\right\}}\sigma_n^2(t).$$

$$r_i(n)=\frac{1}{L}\sum_{k=0}^{L-1}S_{si}(e^{j2\pi k/L})e^{j2\pi kn/L}$$

$$R=\begin{bmatrix}\cos\theta & 0 & -\sin\theta\\ 0 & 1 & 0\\ \sin\theta & 0 & \cos\theta\end{bmatrix}$$

(a) Correct recognition results

$$\mu_\ell=\sum_{k=-\infty}^{\infty}{}^{m}g_{\ell-k}\eta_k,$$

$$\max_i\{\mu_i\}=\max_i\{$$

Original images

$$\mu l=k\sum_{=-\infty}^{\infty}mgl-k\eta k$$

$$m\overset{a}{i}x\{\mu_i\}=m\overset{a}{i}x\{$$

Reproduced images by TEX

(b) Errors in recognition results

図 7　認識結果の例
Fig. 7　Some examples of recognition results.

いる．構造が正しく認識されたものの例を図7(a)に，間違ったものを図7(b)に示している．これらの認識結果は，木構造の表現を，自動的にTEXフォーマットに変換するフィルタを通して再現したものである．また認識時間はいずれの場合も2，3秒以内であった．

数式の構造認識では，文字認識のように大量の認識実験を行い，認識率を定量的に評価することは困難である．このため特に認識率は求めていないが，図7から明らかなように，各種の数式構造が本手法により，正しく認識されていることがわかる．誤認識例の上段は添字式の判定誤り，下段は周辺分布による水平方向の分割の段階で，"a" と "i" を同時に切り出したためである．他の代表的な誤りの多くは，記号同士の接触によるものであった．

4. むすび

本論文では数式の構造認識を行う場合，水平・垂直方向の周辺分布による再帰的分割により，数式の大まかな構造を取り出す手法が有効であることを示した．今後の課題としては，今回は全く考慮していない記号同士の接触の処理，周辺分布による不適当な分割で生じる誤りの回復，より広範囲の数式を高精度で認識するために，数式の構文知識を部分的に利用する手法等を検討する予定である．

文　献

(1) Fu K. S. : "Syntactic Method in Pattern Recognition", Academic Press, pp. 245-252 (1974).

(2) Wang Z. and Faure C. : "Structural Analysis of Handwritten Mathematical Expressions", Proc. ICPR, pp. 32-34 (1988).

(3) Chou P. A. : "Recognition of Equations Using a Two-Dimensional Stochastic Context-Free Grammar", SPIE, 1199, pp. 852-863 (1989).

(4) 岡本正行，西沢　一："数式を含む英文論文誌読取りシステムの試作と実験"，信学技報，PRU88-158 (1989-03).

(5) Okamoto M. and Miyazawa A. : "An Experimental Implementation of a Document Recognition System for Papers Containing Mathematical Expressions", Structured Document Image Analysis, Springer-Verlag, pp. 36-53 (1992).

(6) Okamoto M. and Miao B. : "Recognition of Mathematical Expressions by Using the Layout Structures of Symbols", Proc. ICDAR'91, pp. 242-250 (1991).

(平成6年6月30日受付)